Interactive comment on “Towards an improved description of ocean uncertainties: effect of local anamorphic transformations on spatial correlations” by J.-M. Brankart et al.

J.-M. Brankart et al.
Jean-Michel.Brankart@hmg.inpg.fr
Received and published: 16 January 2012

We thank the reviewer for his/her careful reading of our paper, for his/her acknowledgment of the general quality of the manuscript, and for his/her remarks that will help improving the clarity of the mathematical background. We did our best to take them into account as explained below.

We agree with the reviewer that it is usually better if the mathematical reasons explaining the effect that is observed are explored before the applications. This is why we have added a new section 2.4, which provides the required theoretical background (see below). However, it is also important not to forget that illustrating the importance of this effect is not the only originality of this paper (as pointed out by reviewer 2), which is also written to show that an accurate approximation for the anamorphic transformations (providing a general non-Gaussian description of the marginal distribution for each random variable) can be obtained using a technically simple and efficient algorithm. Moreover, in our case, the theoretical basis for the effect already exists (as also pointed out by reviewer 2) and we used five examples to show how important it is in various ocean applications:

1. First of all, as mentioned in the introduction, we already studied the effect of anamorphic transformations on correlations in a previous paper by Béal et al. (2010). In that paper, we already explained the mathematical reason for which anamorphic transformations can lead to a better description of correlation (i.e. the replacement of the linear correlation coefficient by a nonparametric measure of correlation like rank correlation, see below). We also presented a lot of examples (using scatterplots), which allowed us to discriminate the situations in which (i) the Gaussian assumption is sufficient, (ii) anamorphic transformations improve the description of the data, and (iii) anamorphic transformations do not help (even if they never introduce spurious correlations, and almost never remove meaningful correlations); The purpose of the present paper is then to illustrate the same effect on spatial correlations (not shown in Béal et al., 2010).

2. Second, it is not really exact to say that we do not explain the effect of the transformation on the correlations. It is not done in section 2, but in the examples. First, in the description of Fig. 5:

“This means that the MLD response to Gaussian parameter perturbations is not Gaussian, as illustrated in Fig. 5 (left panel) by a scatterplot of MLD vs SST at 114°W 0°N. As a consequence, the joint distribution of MLD and SST cannot be bi-Gaussian, as visually obvious from the clear nonlinearity of the regression
line (i.e. the line of maximum MLD probability for every given SST). In the transformed variables (Fig. 5, right panel), even if the marginal distribution for each variable is now close to Gaussian (by construction), the joint distribution is still not bi-Gaussian (larger MLD dispersion for small SST than for large SST). But at least the regression line is now close to linear, with the direct consequence of increasing the linear correlation coefficient. This phenomenon explains why the spatial correlation structure can only be improved by consistent local anamorphic transformations.”

And second, in the description of Fig. 6:

“Going to a nonlinear measure of correlation (like the rank correlation, in the middle panels of Fig. 6) is only useful if the transformation can help linearizing the regression line between the two random variables (as illustrated in Fig. 5). The rank correlation was indeed introduced by Spearman (as explained by Von Mises, 1964) to produce this effect and thus to go beyond the linear correlation coefficient (of Pearson), as a measure of the (nonlinear) dependency between random variables. Furthermore, since the linear correlation structure after a local Gaussian anamorphosis is very similar to rank correlation (compare right and middle panels in Fig. 6), this explains why the correlation radius is generally increased by the transformation.”

The same explanations apply to all following examples, which is why it is also summarized in the conclusion:

“These effects may be understood by observing that the linear correlation coefficient (Pearson) between the transformed variables corresponds to a nonlinear measure of correlation between the original variables, which is very similar to the rank correlation (Spearman).”

3. And third, there is an abundant statistical literature discussing the advantages of nonparametric correlations (like Spearman’s rank correlation) as compared to the linear correlation coefficient (Pearson). In particular, nonparametric correlations are (a) more adequate to see a nonlinear dependence between random variables, and (b) more robust to the presence of outliers in the data. This can be illustrated by the famous example of the Anscombe’s quartet (Anscombe, 1973):

http://en.wikipedia.org/wiki/Anscombe%27s_quartet

showing that the linear correlation coefficient is unable to see the perfect (nonlinear) dependence between the random variables in example (2), and is very sensitive to the presence of outliers in examples (3) and (4). (Anscombe, 1973 already stated that “case (2) can sometimes be brought back to case (1) by transforming the x-scale or the y-scale or both”, which is exactly what is done with anamorphic transformations of the variables.) The advantage of using nonparametric correlations in such cases is also explained in many textbooks, for instance, in wikipedia:

http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

where the basic phenomenon is well illustrated by scatterplots, and many references are given. It is particularly clearly and briefly summarized in Numerical Recipes (Press et al., 2004):

“We could construct some rather artificial examples where a correlation could be detected parametrically (e.g. in the linear correlation coefficient r), but could not be detected nonparametrically. Such examples are very rare in real life, however, and the slight loss of information in ranking is a small price to pay for a very major advantage: When a correlation is demonstrated to be present nonparametrically, then it is really there! (That is, to a certainty level that depends on the significance chosen.) Nonparametric correlation is more robust than linear correlation, more
resistant to unplanned defects in the data, in the same sort of sense that the median is more robust than the mean.”

Moreover, these explanations from the statistical literature are equivalent to the reasons given in the field of Geostatistics for the effect of anamorphic transformations on the correlation structure (as pointed out by the second review).

Consequently, what is shown in the paper already closely corresponds to what is suggested by the reviewer, i.e. use a selection of examples to illustrate a result that is generally valid (and more specifically, in our case, show the importance of this effect in many ocean applications).

Nevertheless, it is true that the clarity of the manuscript could be improved by giving the theoretical basis in section 2, before going to the examples, instead of explaining things step by step as the examples become more and more complicated. This is why we have added a new section 2.4, which provides a brief summary of the theoretical background, together with references in which the reader can find more comprehensive explanations:

2.4 Effect on correlations

However, since the examples given in the following sections are mainly dedicated to illustrate the effect of anamorphic transformations on spatial correlations, it is certainly useful to provide first a summary of the theoretical background explaining the effect that can be expected. For that purpose, we assume that we have two non-Gaussian random variables \( X_1 \) and \( X_2 \) (with marginal cdfs \( F_1 \) and \( F_2 \)) that have been transformed into the Gaussian variables \( Z_1 \) and \( Z_2 \) (with the same cdf \( G \)). First of all, it is important to remember that, since the transformations are invertible, there is no loss of information induced by the anamorphosis, and the statistical dependence (in a general sense) between the random variables remains unchanged, i.e. the reduction of entropy gained from the knowledge of the other variable (i.e. the mutual information \( I \)) remains the same:

\[
I(X_1, X_2) = H(X_2) - H(X_2|X_1) = H(Z_2) - H(Z_2|Z_1) = I(Z_1, Z_2)
\] (1)

which can easily be verified by introducing the change of variables in the definition of entropy \([H(X_2)]\) and conditional entropy \([H(X_2|X_1)]\). Consequently, it is only the effect of anamorphic transformations on linear correlations that we are going to investigate, since this is the only kind of correlation that can be described by a Gaussian model.

A first insight into this problem can easily be obtained by remarking that, if there exists separate bijective transformations for \( X_1 \) and \( X_2 \) transforming their joint non-Gaussian distribution into a bi-Gaussian distribution for \( Z_1 \) and \( Z_2 \), then the anamorphic transformation given by Eq. (1) [in the paper] provides the required transformations. This is obvious since the marginal pdfs of a bi-Gaussian distribution are both Gaussian, and the only backward anamorphosis (except for any unimportant additional linear change of variable) transforming the Gaussian marginal pdf for \( Z_1 \) and \( Z_2 \) into the right marginal pdfs for \( X_1 \) and \( X_2 \) is the one given by Eq. (1). In this ideal case, the mutual information is related to the linear correlation coefficient \( \rho_{Z_1 Z_2} \) between the transformed variables (e.g. Cover and Thomas, 2006) by:

\[
I(X_1, X_2) = I(Z_1, Z_2) = -\frac{1}{2} \ln(1 - \rho_{Z_1 Z_2}^2)
\] (2)

As a direct corollary, we can see that, if the variable \( X_1 \) and \( X_2 \) are tightly correlated along a monotonic nonlinear curve (i.e. the ideal situation to estimate \( X_2 \) from an observation of \( X_1 \), but in which linear estimation methods can be very inaccurate), then the anamorphic transformation will transform this curve into a straight line (so that the two marginal pdfs can be simultaneously Gaussian). In this case, the nonlinear depen-
The oldest and most simple example of a nonparametric measure of correlation is the rank correlation (Spearman, 1904; Kendall, 1962), which is defined as the linear correlation between the rank of each member in the ensemble. Hence, this corresponds to a linear correlation between uniform sets of integers between 1 and \( m \), which is thus close to computing a linear correlation after a uniform anamorphosis (i.e., with a uniform target pdf), instead of a Gaussian anamorphosis. (This is only approximate because, unlike uniform anamorphosis, the computation of the rank is not invertible, so that there is a small loss of information in the operation.) The close similarity between the rank correlation between \( X_1 \) and \( X_2 \) and the linear correlation between \( Z_1 \) and \( Z_2 \) was already discussed in Béal et al. (2010), and it is further illustrated here in the example of section 4 (Fig.6). And it is the use of such a nonparametric measure of correlation between \( X_1 \) and \( X_2 \) (i.e., the linear correlation coefficient \( \rho_{Z_1Z_2} \) between the transformed variables \( Z_1 \) and \( Z_2 \)) instead of the linear correlation coefficient \( \rho_{X_1X_2} \), that is the fundamental reason explaining the improvement of the correlation structure that is observed in the rest of this paper, and that was also observed in other applications of anamorphosis in Geostatistics (e.g., Chilès and Delfiner, 1999).

Other remarks:

1. We do not agree with the statement that “the authors rely on the assumption that only a Gaussian description of uncertainties is reliable”, since we explain throughout the paper that anamorphic transformations (if diagnosed from the ensemble) provide a general non-Gaussian description of the marginal distributions, and since it is explained in the introduction why it may often be a good practical compromise:

“However, even if an explicit stochastic modelling is used to solve a practical problem, there is often a strong temptation (in large size applications) to sim-
plify the result using a Gaussian model, because it is much more efficient (i) to
describe the uncertainties (by the mean and covariance), and (ii) to assimilate
observations (using linear update formulas, as in the ensemble Kalman filter,
see Evensen and van Leeuwen, 1996). Without a prior assumption about the
shape of the probability distribution, large size problems are indeed very com-
plex in general (van Leeuwen, 2009; Bocquet et al., 2010), mainly because the
size of the sample that is required to identify a general multivariate distribution
increases exponentially with the number of dimensions (curse of dimensionality).

To circumvent this difficulty, one possible simplification is to look for univariate
nonlinear changes of variables (anamorphosis transformations) transforming the
marginal distribution of each random variable into a Gaussian distribution. One-
dimensional probability distributions can indeed be identified with a much smaller
sample, and it may well happen that such a separate transformation for each
random variable also helps improving the Gaussianity of their joint distribution
(although this needs to be checked in every practical application).”

The first paragraph of the conclusion looks also quite clear about that:

“Many kinds of ocean uncertainties cannot be accurately described using a Gaus-
sian model. This is particularly obvious in the examples of ecosystem uncertain-
ties (in sections 4, 5 and 7) and sea ice uncertainties (in section 6), although this
may also be true for ocean dynamics uncertainties (as in the mixed layer depth
example in section 3). (…) Nevertheless, even with the available ensemble (a
few hundred members in all examples described in the paper), it is certainly pos-
sible to go beyond the Gaussian assumption in the description of the marginal
distribution for any individual random variable (…). In this paper, we suggested
that a very significant improvement can already be obtained with a very simple
non-Gaussian description of the marginal distributions (histograms), based on a
few quantiles of the ensemble (typically deciles, as in our examples). (…)”

2. It is incorrect to say that “the anamorphosis transformation is performed inde-
pendently for each single grid point”, or that they are “different and unknown
functions”, because they are diagnosed from the ensemble to transform (approx-
imately) each marginal pdf into a Gaussian pdf. Thus, if the random variables at
every model grid point are not independent, then the transformation are also not
independent. On the contrary, the transformations are exactly what is needed to
transform a linear correlation into a nonparametric correlation (resembling rank
correlation, see above).

3. About the large correlation between the Loop current and the Western coast of
the Gulf of Mexico, we agree that they cannot be expected to represent real
model errors. The large correlations are due to the very simplistic assumption
that is made to generate the ensemble (constant parameter perturbations over
the whole Gulf of Mexico). Our purpose is here to evaluate the effect of anamor-
phosis transformations on correlations, not to discuss the validity of the ensemble
to represent actual model errors. See our answer to the minor comment 2 of re-
viewer 2 for more details, and for the clarification that we have included in the
paper.

Additional references

Anscombe F. J.: Graphs in Statistical Analysis, American Statistician, 27(1), 17–21,


Corder G. W., and Foreman D. I.: Nonparametric Statistics for Non-Statisticians: A


Kendall M. G.: Rank correlation methods, Griffin, 1962.


Interactive comment on Ocean Sci. Discuss., 8, 2147, 2011.